John von Neumann Institute for Computing

Statistical Mechanics of Protein Folding: Some Outstanding Problems

Alexander Grosberg

published in

Computational Soft Matter: From Synthetic Polymers to Proteins, Lecture Notes,
Norbert Attig, Kurt Binder, Helmut Grubmüller, Kurt Kremer (Eds.),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **23***, ISBN 3-00-012641-4, pp. 375-400, 2004.*

http://www.fz-juelich.de/nic-series/volume23

Statistical Mechanics of Protein Folding: Some Outstanding Problems

A. Grosberg

Department of Physics, University of Minnesota, Minneapolis, MN 55455, U.S.A.
and
Institute of Biochemical Physics, Russian Academy of Sciences, Moscow 117977, Russia
E-mail: grosberg@physics.umn.edu

We consider broken ergodicity, memorization of random choice, and spontaneous symmetry breaking in both sequence space in the evolution and in conformation space in connection with folding. We pay major attention to the topology of conformation space, its connectivity, and the determination of both the order parameters and the (free) energy landscapes.

1 Introduction

It is useful to start the introduction with the recommendation to the reader. If one wants to study proteins, then there is a very good book to jump-start the process, it is "Lectures on Protein Physics," by Alexei Finkelstein and late Oleg Ptitsyn[1].

1.1 Basic Terminology

Proteins are heteropolymer molecules, consisting of 20 different monomeric units, called amino acid residues, or simply amino acids. These units are connected in a chain-like topology, such that the main chain of the polymer, called polypeptide chain, is universal, does not depend on the specific types of the amino acids, while different amino acids form different side groups at every chain monomer.

Proteins are characterized by their primary, secondary, ternary and quaternary structure. The primary structure is defined as the sequence of amino acids along the polymer (polypeptide) chain. The fact that polypeptide chain is regular, sequence independent, gives rise to the formation of certain regular conformational elements, such as helices or sheets (called, since Linus Pauling, α-helices and β-sheets). These elements are usually referred to as secondary structure. However, in order to perform its biochemical function, the overwhelming majority of the proteins acquires a complex three-dimensional structure during the protein folding process. Protein folding is the process in which molecule self-organizes into a native three-dimensional structure, called ternary structure. Finally, the quaternary structure is the arrangement of the constituting ternary structure domains, belonging either to the same or to different polymer chains.

The difference between secondary and ternary structure frequently seems confusing for the beginners. One aspect of the difference, which appears purely quantitative, if not vague, is that secondary structure, such as spiral-like α-helix, involves relatively short pieces of the chain, while folding of the ternary structure involves spatial scales of the entire molecule. Another, possibly more clear aspect, is that secondary structure motives are determined by the geometry (or stereochemistry) of the main polymer chain. For instance,

in α-helix, there are hydrogen bonds between monomers with numbers k and $k+3$ along the chain (for some series of subsequent numbers k). Historically, people thought that energetics of the secondary structure elements is also due to the interactions between the main chain atoms, or, better to say, due to the competition between hydrogen bonds of the (main chain) peptide groups to each other or to surrounding water. In fact, bulky side groups, although located outside α-helix, or above and below the β-sheet, can be in proximity to each other and, therefore, can affect energy and stability of the corresponding secondary structure elements. In other words, the stability of secondary structure is strongly affected by the ternary structure, and vice versa. Nevertheless, the distinction between secondary and ternary structure is fairly sharp.

The protein folding process, during which the heteropolymer chain reaches its folded, native structure, is one of the central, but not yet completely understood problems of the molecular biology. This lecture gives an introduction to some of the most pressing issues of protein folding from the viewpoint of the theoretical physics (or, it maybe better to say, from the personal subjective viewpoint of the present author).

1.2 Folding and Re-Naturation

To describe the "creative" process of protein folding, it is convenient to start with the "destructive" process of protein denaturation. Denaturation is what happens when one boils an egg. In a better controlled experiment, one can take a very dilute solution of native proteins, where every protein molecule is very far and independent from any other, and thus acts independently of others (unlike in the egg). Then, when the solution is heated up, something happens. For instance, if the protein is an enzyme, it keeps working as a catalyst up until certain temperature, called denaturation point, above which no appreciable catalytic activity is observed.

The story of denaturation studies is a long and interesting one. For instance, it was found that temperature of denaturation for every protein is not very far above the temperature at which host cell normally lives and functions: proteins from bacteria living on glaciers denature at about $20^\circ\mathrm{C}$, while organisms living in the hot water of geiser have their proteins stable up until $140^\circ\mathrm{C}$ or so. Another exciting observation is that denaturation is a cooperative all-or-none transition, it is a finite system counterpart of the first order phase transition[2]. Denaturation transition is accompanied by a significant latent heat, usually about $7 \times 10^{-21}\mathrm{J/monomer}$. Per unit mass, it is about the same order of magnitude as the latent heat of melting of usual molecular crystals, such as naphtalene. Finally, denaturation can be caused not only by heating, but also in many cases by cooling down ("cold denaturation"), as well as by changing pH, salinity, etc, and, most easily, by adding to the solution some "denaturing agents," most of which disrupt hydrophobic interactions (such as urea).

What happens to the molecule when protein denatures? In many cases, although not in all cases, it is globule-coil transition: native globule opens up and becomes essentially a Gaussian coil, in which volume interactions between aminoacids are of minor importance. In some other cases, native globule gets transformed into another globular form, called molten globule, which does not have so uniquely determined ternary structure. Once again, we should refer the reader to the book[1] for further details.

In fact, when protein denatures, one should be wondering if some irreversible changes occurred, such as breaking of covalent bonds connecting protein polymer chain together.

Indeed, hard boiled egg is certainly irreversibly changed from the fresh, no chicken can be produced from it by any manipulations. Is the same true for proteins denatured in the dilute solution? It was a major discovery by C. Anfinsen about 50 years ago that gently denatured proteins can be returned back to their native states, their enzymatic activity can be restored, and, therefore, nothing irreversible happens to them during denaturation. The phenomenon of return to the native state is called re-naturation. After this was first established in classical experiments on ribonuclease by C. Anfinsen[3], renaturation was observed with many proteins.

Anfinsen's result means, first of all, that all the information necessary to form the native ternary structure is contained in the molecule itself, because renaturation is observed in a very dilute solution, where protein molecules do not interact. Therefore, all the necessary information is contained in the primary sequence. In Anfinsen's own words in his Nobel acceptance speech in 1972[4]

> "The native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment."

As we already said, not all of the proteins renature starting from completely open coil-type state. Furthermore, not all of the proteins renature at all. It is however most important to realize that there **exist** many proteins which do not require anything, like any external help or assistance, to fold, their folding is an example of true *self*-organization. A lot of proteins will correctly fold into their native state in very dilute water-based solutions, provided only with appropriate pH and salt concentration. There are some other proteins which do require assistance to fold; they only fold in the presence of chaperons. Note also that not all of the proteins function in the folded globular state. Some, such as, e.g., collagen function as fibrils. Others are attached to membranes. These lectures do not discuss these cases. However important biologically, all these aspects are of secondary relevance to the physics approach to the subject of folding. Since there are proteins which do fold without any help, it is clear that they should be the first subject of attention and they should be understood first.

As Anfinsen experiment suggests beyond a doubt, native state of a foldable protein is stable and thus represents a minimum of free energy. This does not mean it is necessarily the global minimum. In other words, Anfinsen's experiment leaves open the possibility that native state is just a local free energy minimum and that there is one or more even more stable states, which correspond to deeper minima. This possibility is very likely realized in some cases (e.g., prion proteins), when both correctly folded and some misfolded states are realized under different circumstances. These misfolded states are very important, because they are responsible for a number of diseases, like, e.g. some illnesses of the nervous system ("mad cow disease"). This already suggests the importance of both thermodynamic (equilibrium) and kinetic aspects of folding, because folded state may be either thermodynamically most stable (global minimum), or thermodynamically metastable (local minimum), but kinetically most easily accessible.

1.3 Quenched, but not Disordered

It is useful to emphasize from the very beginning that the physics of protein conformations, including the folding phenomenon, is closely related to the branch of physics which

is called physics of disordered (or, sometimes, "dirty") systems, such as doped semi-conductors. The relation between proteins and disordered systems may appear bizarre and unfounded. Indeed, proteins are very pure: each protein molecule is an exact copy of every other molecule of the same protein, they have identical molecular mass, identical number of monomers in the chain and, moreover, identical sequences of links along the chain. The confusion here is entirely due to the historically accepted terminology, which for the regular systems, such as semi-conductors, concentrates on the aspect of disorder. Indeed, donor and acceptor atoms in a semi-conductor are positioned randomly. These samples are manufactured from a melt, a liquid mixture of different atoms, and when melt is quenched and frozen, the minority atoms in the host lattice are essentially fixed at the random places where they happen to be at the moment of quench.

Similar situation is known in the materials called spin glasses. Those are the dilute solutions of paramagnetic atoms in the non-magnetic matrix. Once again, if we start from a melt, and then quench it, then paramagnetic centers are caught in the random places where they happen to be at the moment of quench.

When such systems are called disordered, it is meant to emphasize that although they are solids, they have no translational or any other simple order in their structures. In protein context, this aspect of disorder is neither interesting nor productive. More important is the aspect that the structure is quenched. That means, for the spin glass example, that the spatial positions of the paramagnetic centers are fixed and, once the sample is prepared, they do not change. This is interesting, because every paramagnetic atom has obviously a spin, and exchange interactions between these spins is determined by the distance between them in space. Thus, in essence, when spin glass sample is quenched, what happens is we have a system of spins, s_i, $i = 1, \ldots, N$, with interaction between any two spins s_i and s_j, $J_{ij}s_is_j$, where matrix J_{ij} is quenched. Thus, physics of spin glasses is the statistical mechanics of the spin system which is pretty much like the regular Ising system, except interactions matrix is not regular, but quenched.

In proteins, sequence, or primary structure is under strict genetic control. This means, every protein molecule is prepared by the protein synthesis cell machinery under DNA and RNA guidance, such that all copies of the same protein have identical sequences. When protein is prepared, its sequence does not change during folding or any other conformational transformations. In this sense, proteins are the systems with quenched sequences. In this respect, whether we call proteins ordered or disordered is more emotional than scientific issue, the only thing that is important is that they are quenched.

This said, it is important to understand where do these sequences come from. In fact, they are the product of biological evolution.

1.4 Quenching and Self-Averaging

In order to understand the above concepts, let us consider a "photo" of an Ising spin system, where the spins are oriented randomly. The randomness can be caused by two reasons: either we are above the critical temperature in the uniform Ising system with ordered interactions, or we deal with irregular system with disordered interactions.

In the first case the spin configuration changes in time, and it is a realization of the thermal noise. In the second case the spin configuration also changes in time, but does so on the background of a "prepared" interaction pattern which does not change in time. How-

ever, one "photo" cannot make distinction between the cases. The first is called annealed, while the second quenched disorder.

In spin glasses, both kinds of disorder are present. These materials are alloys, where paramagnetic atoms are placed in random positions, and their interaction is ferromagnetic or antiferromagnetic depending on their relative distance. The spin degrees of freedom can freely rotate or flip, thus they are annealed, while the position of the paramagnetic atoms and, therefore, interaction of spins is fixed, thus they are quenched. The Hamiltonian of such a system has the form of $H = \sum_{ij} J_{ij} s_i s_j$, where i and j label the paramagnetic centers. While calculating the partition function of such a system the terms J_{ij} describing the quenched interaction has to be kept constant, and we have to sum over the annealed spin configurations only.

Similarly, the sequence of a polymer can be considered as quenched (neither the order, nor the character of the amino acids are not changing after a protein is formed), while its conformation (that can change with thermal motion) an annealed variable.

Usually in physics, quenched variables are disordered, like, e.g., in a spin glass. In proteins, it is meaningless to say that the sequence is disordered, but it is certainly quenched.

What complicates calculation of the partition functions $Z(S)$ is the fact that this calculation, in principle, must be done separately for every sequence S, and there are many of them. However, if the partition function is known, all the thermodynamic quantities of the protein can be calculated. It turns out, that two situations are possible in this calculation. They are called self-averaging and non-self-averaging, respectively.

A quantity is called self-averaging when its value is approximately the same for each realization of the quenched variable (eg. samples of spin glasses). Additive quantities, like the free energy, usually are self-averaging, but the multiplicative ones, like the partition function are not self-averaging for sure.

If a quantity is self-averaging, then it is useful to compute its average over the disorder (over sequences), because the result for (almost) every particular sequence will be close to this average. This is the case with free energy, so $\langle F(S) \rangle$ is a reasonable useful quantity. This approach makes no sense for the non-self-averaging variables, and so $\langle Z(S) \rangle$ is not a good characteristic.

The reason why $F(S)$ is self averaging and why $Z(S)$ is not can be explained a little more formally, as follows. The free energy is an additive, or extensive, variable. We can imagine our system as consisting of many sub-systems, each of which is still macroscopic. Then, the free energy is the sum of free energies of these subsystems. Each of the subsystems has its own realization of the disorder, and so their free energies are statistically independent from each other. Then, central limit theorem applies, and says that the free energy of the whole system is narrowly (Gaussian) distributed, with the width that goes down as the system size, or the number of monomers, N, increases. That precisely means that $F(S)$ is self-averaging.

Why does not this argument apply to $Z(S)$, where does it fail? It fails because Z is not an additive quantity. Indeed, as $Z(S) = \exp\left(-F(S)/k_B T\right)$, this quantity is exponential in N. That means, if one particular sequence S has free energy lower than all other sequences, then its partition function is exponentially higher than partition functions of all other sequences. Such exponentially large quantity may (and does!) dominate the average $\langle Z(S) \rangle$ even if this particular sequence has very low probability, such as, e.g., a homopolymer or exactly alternating sequence such as $10101010101\ldots$.

1.5 Plan

In this lecture, we shall discuss basic issues in the statistical mechanics modeling of proteins, which involves both sequence evolution and conformation dynamics.

2 Sequences

2.1 Diffusion in the Sequence Space

Evolution of sequences is itself an exciting problem with a large statistical mechanics component. Needless saying, it has got particularly exciting now, with plenty of uncovered experimental (statistical) material. The general understanding of the area is impossible without the realization of the basic fact: evolution of sequences is about as old as the Universe itself. Indeed, according to the modern estimates, the age of the Universe since the Big Bang is about 15 billion years, the age of the Earth is about 4.5 billion years, and the oldest clear signs of life on the Earth are about 3.8 billion years old. So, all these numbers are fairly close. Thus, speaking about evolution of sequences, we should have in mind an event which happened once and did not have time to be repeated to explore different possibilities.

Let us look at evolution mathematically, considering it as some sort of diffusion in the sequence space. This sequence space is obviously a discrete one: if we have, say, Q possibilities for every monomer (Q letters in the alphabet, $Q = 20$ for real proteins), then we only have a finite number of sequences Q^N, where N is the maximal acceptable length. This space is big, because the number Q^N is astronomical. In fact, this number is so large that, for instance, H. Fraunfelder[5] suggests to call it a "biological number," where biological numbers dwarf astronomical ones. Two ingredients are vital to imagine diffusion in such space: first, we should allow some dynamics, which means there should be some mechanism allowing sequences to change from time to time producing some other sequences; second, the system must be able to decide which of the newly produced sequences are good and which are bad, which means there should be a feedback.

A few comments are in order. In real biology, the mechanism of sequence change is incredibly complex, it involves mutations in DNA, biosynthesis of proteins based on mutated DNA, their folding, and inheritance of the mutated DNA to new generations of organisms. By the way, real mutations in DNA are not only point mutations, there are things as sophisticated as swapping of long blocks etc.

Let us simplify this as far as possible. Namely, let us imagine that the time is discrete, and pretend that at every time step, with every "click of the evolutionary clock" every sequence, S, can jump and transform into some other sequence, S'. In reality, these time steps are perhaps close to the life time of one cell before it produces the offspring.

This gives rise to a convenient mathematical scheme. Let us imagine every possible protein sequence as a dot and let us connect two dots by a line if they can be transformed into one another in one time step. We call these lines "mutations," although they are not real mutations, real mutations occur in DNA, and we speak of the effective "mutations," which are the jumps in the space of protein sequences. Thus, with sequences= represented as nodes and connected by bonds representing "mutations," our sequence space is now presented as a graph.

Now we return to the question of the feedback. Arriving at every sequence, or every node on the sequence space graph, our system has to be able to decide whether this sequence is good and should be multiplied, or it is bad and should be eliminated. Phenomenologically, we should imagine that there is some function U which is defined for every sequence, $U(S)$, and which measures fitness of this sequence. By analogy with energy, it is common to define $U(S)$ such that the better fitness corresponds to the lower values of $U(S)$.

In real biology, once again, this is incredibly complex. The "test" involves folding of a protein with the new sequence, its function, and whether it is appropriate for the cell and the organism. In many cases, perhaps in the majority of cases, the test is an outright failure: either the polymer does not fold, or does not function well enough, or does some other nasty things, such as aggregates with something else. Rather rare steps lead to something very good.

According to Kimura and his neutral theory of evolution, there are also great many steps which are neutral, they lead to the sequences which under current conditions are neither better nor worse than existing ones. This neutral theory suggests that diffusion over neutral regions of the sequence space creates a pool of sequences to choose from when the conditions change.

Now, we can try to formulate the mathematical model of diffusion in the sequence space. One way to do this is to use the Metropolis Monte Carlo method, considering fitness as effective energy (divided by $k_B T$). During the "evolution" represented by the Monte Carlo dynamics of the Metropolis algorithm, a biased diffusion will take place in the sequence space. The equations for the "flux" from site S to S' or back are such that the system generally moves in the direction of decreasing U, however, with some small but non-zero probability it can also move in the opposite direction, which allows overcoming the local barriers and prevents the system from falling down in a local trap. More comments about Metropolis method will be given later, when we shall use it to study diffusion in conformation space.

2.2 Spontaneous Symmetry Breaking and Memorization of The Random Choice

The Metropolis scheme is not the only possible one to describe biased diffusion in the sequence space, nor is it necessarily the best one. It is brought here as just an example. One can very easily argue about lots of important factors which can be included only in a doubtful way. Perhaps the most important is the fact that proteins interact strongly, they interact in both physical and information sense, which means that the performance, or fitness of a given sequence, depends strongly on the set of other sequences currently in the pool. However correct are all the critical remarks, there is one general conclusion which is robust, universal and fully insensitive to the details: it is that evolution did not have time to visit all sites in sequence space, the evolution had time to try only a very small fraction of the sequences.

Indeed, it is a very general property of diffusion type equations in a restricted domain to exhibit two distinct regimes, which correspond to early time and late time asymptotics. The late time asymptotics is characterized by exponential approach of the distribution to the "ground state" eigenfunction of the corresponding operator. The existence of such eigenfunction is guaranteed when the domain is restricted. This late time asymptotics

corresponds to exhaustive (or ergodic) sampling of all possibilities. This is what happens in the thermodynamic equilibrium. On the other hand, the early time asymptotics corresponds to the initial spread out of the diffusing cloud (or probability distribution), in this regime overall size and shape of the domain is not relevant. In this sense, evolutionary diffusion in the sequence space is currently in the very early stage, it is very far from exhaustively sampling all possible sequences. This most fundamental fact easily follows from the very simple estimate. Total number of sequences of the length, say, $N = 200$, is $20^{200} \approx 10^{260}$. This is a "biologically large" number[5]. It is easy to establish that even if we use all the mass of the Earth (6×10^{24}kg) to make proteins (mass about 3.4×10^{-24}kg each), and will change all the sequences every nanosecond, even then the total number of tested sequences can reach "only" about 10^{74}, which is still nowhere near the 10^{260} required. Therefore, most of the possible protein sequences have never been "tested": we are in the short-time limit. In particular, we cannot say that current proteins are the best possible ones, because Mother Nature did not try most of others...

The fact that all sequences could not possibly be tested indicates that the existing choice of sequences has a significant element of randomness to it. As Oleg Ptitsyn expressed it once[6], proteins are slightly edited random copolymers. Edited - because they did go through some evolutionary selection; but edited only slightly - because we are still in the early time limit in terms of diffusion in the sequence space.

3 Conformations, Conformation Space, and the Levinthal Paradox

3.1 Conformation Space: Lattice Polymer

Computationally most tractable polymer model is the lattice one. It represents the polymer as a self avoiding walk on a (cubic) lattice. An appropriate set of local moves is usually defined to allow the dynamics realized as a succession of these moves. In the most common Stockmayer-Verdier model[7], these moves include end flip, corner flip, and crankshaft flip. The only requirement for the set of local moves is that they have to provide for ergodicity: every conformation should be possible to transform into every other conformation. Importantly, ergodicity does not require that the path from one conformation to another be short or simple, it only requires that such path exists.

We now have to make one step further in the direction of abstraction and ask: what is the space of conformations for such lattice polymer? This space is in fact a graph[8,9]. Each vertex of the conformational graph represents one particular conformation. Two vertices of the graph are connected by a bond if and only if the corresponding conformations can be transformed into one another by a *single* elementary move. When all conformations are included, this graph is connected, which is to say that the system is ergodic. In general, this graph, or certain regions of it, may have non-trivial fractal dimension[10]. In addition, certain parts of conformation graph are shown to be of a small world network type[11]. Finally, if we restrict consideration with only compact conformations, then ergodicity may be broken[12]; controlling, say, maximal gyration radius of the polymer in real space we can observe percolation transition in the graph of conformations.

3.2 Conformation Space: Off-Lattice Polymer

In the simplest off-lattice model, the chain conformation is fully described by the set of monomer position vectors, $\boldsymbol{r}_i$. Naively, this suggests that the space of conformations for the N-mer is a $3N$-dimensional Euclidean space. This is wrong. To see why and how this is wrong, it is enough to consider a "polymer" with two monomers ($N = 2$) with fixed bond length: apart from translational freedom of the first monomer, the "space of conformations" is, of course, a sphere - curved surface with non-Euclidean internal geometry. This very simple argument can be pulled quite far to show that conformation space is very much curved for long polymers, and its internal geometry is not at all Euclidean[13]. One should not be distracted by the fact that conformation *can* be fully characterized with the Euclidean position vectors $\boldsymbol{r}_i$ - this only means that the curved space of conformations can be *embedded* in a bigger Euclidean space - just like we usually imagine a sphere in a $3D$ space. This embedding in no way cancels or downplays the importance of the fact that internal geometry of conformation space is not Euclidean, because of course all the dynamical trajectories of the systems are in this space, not in the bigger flat one.

Clearly, the difference between on-lattice and off-lattice cases is neither deep nor important in this context. To understand it, let us begin with neglecting the excluded volume constrains. In this case, lattice model has just 6 (on the cubic lattice in $3D$) possible positions for every bond. Similarly, in an off-lattice model based on rotational isomers, there are a few permissible rotational states for every bond. With 2 possible states for every one of N monomers, the graph of conformations would have been an N-dimensional hypercube; similar figure with 6 or other finite number of vertices along each of the N axis has no special name, so, for the lack of a better word, let us call it a "cube". Of course, imposition of the excluded volume constraints erases many (in fact, even the majority) of the "cube" vertices, the ones which correspond to conformations with overlapping monomers. Nevertheless, the remaining part, which is the real conformations graph, is still a part of a cube, which obviously has nothing to do with the regular lattice - discrete analog of a flat Euclidean space. Similarly, for the off-lattice model with continuous set of rotations, conformation space is compact (in the strict mathematical sense of the word) and curved.

Thus, the stage of the protein folding drama is a compact space, some part of a "cube" or its continuous analog. It is terribly curved and restricted, it has complex topological and fractal properties. How can we understand the motion of a protein in such space?

3.3 Levinthal Paradox

The most pragmatic formulation reduces protein folding problem to that of prediction: knowing the amino-acid sequence of a protein, how to determine its ternary fold? This is potentially a multi-billion dollar issue, as it involves many protein-based drugs. To achieve this goal is very difficult, the problem resists best efforts of best scientists for a few decades now. Why is it so difficult? The nature of difficulty is well explained by the celebrated Levinthal paradox[14]. In the most standard formulation, Levinthal paradox arises from the idea that the time required for a protein molecule to sample all of its conformations is **at least** $M\tau$, where M is the number of distinct conformations, and τ is the time necessary to sample one conformation. Then, the paradox goes, unguided folding into one particular (native) state requires **at least** time of order $M\tau$ which is far too long, because M is astronomically (biologically?[5]) large. Similar to the sequence evolution case described

above, it is easy to establish an emotionally impressive fact, that the age of the Universe is by far insufficient for one typical length ($N = 200$, $M \approx 10^{160}$) protein to sample all possible conformations even if we assume τ as little as $1\mathrm{ns}$.

Of course, the above formulation relies on the discrete conformation space, such as the conformation graph of a lattice polymer. However, the idea remains the same for continuous models as well.

Levinthal paradox was a subject of numerous discussions over the years. One of them was held recently in the Journal of Biomolecular Structure & Dynamics[15]. The present author contribution to this discussion emphasized the role of topology of conformational space. Specifically, the time estimate of the exhaustive search through the conformations depends on how these conformations are mutually arranged, how protein can move from one conformation to another. This aspect was long underappreciated, and deserves more detailed presentation.

3.4 Connectivity of the Conformation Space and Scaling of Levinthal Time

It seems that the idea of this section is well illustrated by the following story told in the author's contribution to the discussion[15].

A long time ago, well before the breakup of Soviet Union, a large biophysics meeting was held in then Soviet Republic of Georgia. The site was a rural place in the center of a famous wine producing region, the month was October, and the major event was the all-Georgian wine testing festival, advertised as a merry traditional peasant holiday. Upon arrival, biophysicists found a large open field in the valley between mountains, covered with dozens of pavilions, each representing a particular village, and each offering for free a glass of young wine. Very soon, the cloud of biophysicists seemed perfectly obeying the diffusion equation, with each individual in the cloud undergoing random walks.

Wine testing continued for an unexpectedly long time. Assuming visit to one pavilion takes time τ, and assuming there were some M pavilions, one could have naively expected that after time close to $M\tau$ the testing would be over. Such expectation proved totally wrong. One possible reason is trivial: wine testers, even if they were to complete the exhaustive testing, are unlikely to realize the completion of the task and to stop at that. But there is also another more interesting reason: the time necessary for the random walk to visit all M sites does not scale as $M\tau$, it can be significantly larger than that, simply because random walk visits some sites great many times before the first visit to some other sites.

Indeed, in terms of wine testing, $M\tau$ is the time required to visit all pavilions in an orderly fashion, one pavilion after another, never returning to the already visited place. Of course, a sober person can do that, but sober model is unrealistic for the wine tester. Equally unrealistic is the model of protein chain dynamics which orderly samples all conformational states, one after another, never returning to the once visited conformation. For all other sampling strategies, the exhaustion time is larger than $M\tau$. Of course, Levinthal did not say that the time of exhaustive conformation sampling (or wine testing) was $M\tau$ - he said it was **at least** $M\tau$; in other words, he said it is $\geq M\tau$. This was sufficient for him to conclude that exhaustive sampling is impossible for realistic N, such as $N = 150$ or 200.

The opposite, and perhaps more realistic, model of protein dynamics (and also of wine

testing) would be purely random walk in the space of conformations. To begin with, suppose it is an unbiased random walk, which means there is no conformation dependent (free) energy landscape involved. How can we estimate the exhaustive sampling time for the unbiased random walk model? Consider first that wine testing pavilions arranged along a line. Then random walk of longevity t brings us as far as about $\sqrt{t/\tau}$, which means we cover all M sites when $\sqrt{t/\tau} \simeq M$, and the time of exhaustive sampling is $t \simeq \tau M^2$. Needless saying, the difference between τM and τM^2 is very significant.

In fact, accurate estimate of exhaustive sampling time by a random walk is not completely trivial. More sophisticated estimate for one dimensional case, which will not be derived here, reads $t \simeq \tau M^2/\ln M$. For the random walk in the space of higher dimension d, the result depends on d. When d crosses over 2, the mechanism of sampling changes, because random walk tends to leave behind large unvisited regions. At $d > 2$, exhaustive sampling is only possible because the overall volume is restricted, and random walk is forced to come back.

What is d in reality is anybody's guess. Please do not forget that d here is the dimension of the abstract space of protein conformations, not the real three-dimensional space.

The result close to $M\tau$ would be correct for the dimension d as high as M; in this case, exhaustion time would have scaled as $\tau M \ln M$. However, this estimate is completely unrealistic, because $d = M$ corresponds to the situation where each conformation (site in conformation space, or wine pavilion) can be equally probably reached from every other conformation in just one step τ. Clearly, real protein chains are nowhere near this extreme.

3.5 Levinthal Exhaustion Time and Folding Time

In fact, the exhaustive sampling time of all conformations is not the quantity of major physical interest. Indeed, we are not interested in trying all conformations, we want to know how long it takes to arrive at the particular native state. More specifically, in the language of random walks, folding time is defined as mean first passage time to the native state starting from an arbitrary open conformation. Here, "native state" usually means one particular native conformation, but, of course, this is just a simplified coarse grained view of things. In fact, native state is also a macrostate, or, in other words, it corresponds to some region, $\mathcal{N}$, in conformation space. This region most likely has a funny non-trivial shape, because being in the native state imposes very strict requirements (say, below 0.1 Å) on some of the coordinates, while some other might be restricted in a much more liberal way (say, 5 Åfor some side groups in the loop regions), or not restricted at all (as, e.g., χ angles of the side groups in the loops). This implies that $\mathcal{N}$ should be thought of as a pancake-shaped or some fractal with large aspect ratio(s).

Furthermore, in reality there are forces and energies involved, which means random walk in conformation space occurs on top of the energy landscape. Since folding is similar to first order phase transition, there should be a transition state, which is similar to a critical nucleus. In this case, the question is how long does it take for a random walk in conformation space to bring the system into some critical region ω. Indeed, in a more physical language, when people speak about nucleation and growth mechanism for the kinetics of the first order phase transition[16], it is assumed that critical nucleus plays the role of a "transition state," and that after having achieved this state the system completes the transition very rapidly. This can be re-phrased by saying that *before* crossing the barrier, or before the

critical nucleus is assembled, the conformational motion can be seen as a largely unbiased random walk in the space of conformations, while the way from the transition to the native state is a bias-dominated walk. Thus, the above mentioned critical region ω is another name for the critical nucleus - it consists of all the microstates which together constitute the transition state ensemble. Thus, our question is how long does it take for the random walk to enter, for the fist time, some region ω starting from a random point in conformation space? To answer this, in addition to the overall conformation space geometry, one has to know more about ω: what is the shape and fractal dimension of ω, what is its boundary, etc. We do not know answers to these pressing questions. Only very crudely, we can imagine that ω, just as $\mathcal{N}$, is in some sense pancake-shaped.

What we do know however is how to formulate the Levinthal estimate of folding time in this language: it is $t \sim \tau \left|\Omega\right| / \left|\omega\right|$, where Ω is the entire conformation space, and $|\ldots|$ means the number of conformations in the domain $\ldots$. This estimate is similar in spirit to the $M\tau$ estimate of exhaustion time, as discussed above, because M, roughly, is $M = |\Omega|$. When random walk terminology is considered, it is obvious that this estimate has nothing to do with the first approach time which we are looking for. Indeed, $\tau \left|\Omega\right| / \left|\omega\right|$ is in fact an average time between two subsequent visits in ω by a very long trajectory assuming that these visits occur rare enough such that roughly ergodic covering of Ω is established between two visits of ω. Thus, in addition to the question of energetic bias towards the native state, an understanding of random walks in conformation space is crucial to the understanding of protein folding.

From our discussion of the exhaustion time, it also follows that the result for the first passage time will very strongly depend on the shape of critical region ω. Very roughly, if we imagine ω as a horizontal pancake, then random walk is only important in vertical direction, and unimportant in horizontal direction. This leads to dramatic reduction of the space dimension in which exhaustive search by a random walk should be performed, thus dramatically reducing the time estimate. This fact was recently emphasized by T.McLeash[17].

3.6 What is Common, and What is Different, Between a Chemical Reaction and a First-Order Phase Transition?

A chemical reaction can be represented in terms of proper reaction coordinates as a movement through a saddle point of the (free)energy surface. The reactants and the products correspond to different local minima. As the temperature increases, the reactants will be able to reach the top of the saddle with an increasing probability, such that overall reaction rate is proportional to $\exp\left[\left(E_{\min} - E_{\text{saddle}}\right)/k_B T\right]$ (there is also a factor determined by the curvature of the free energy surface in the vicinity of the saddle[18]). Temperature dependence of the rate constant is a very steep increasing function, and thus the reaction will proceed with a significant speed only when temperature is not too low.

Somewhat similarly looking but very importantly different probabilities describe the equilibrium properties instead of the rate, this is the probability for the molecule to be in either reactant or a product, $P_{\text{reactant}} + P_{\text{product}} = 1$. The ratio of the probabilities is proportional to $\exp\left[\left(E_{\min 1} - E_{\min 2}\right)/k_B T\right]$. This implies that the thermodynamic (equilibrium) probability to be in either product or reactant state, whichever is higher in energy, is very low at low temperatures, and grows to become $1/2$ at high temperatures. Temperature dependence of this probability is a characteristic sigmoidal shaped curve.

First order phase transition is somewhat similar in the sense that we also have two minima of (free) energy, and they are separated by the barrier. This gives rise to the "switching" behavior somewhat similar to that of the sigmoidal temperature dependence described above. There is, however, an important difference: the width of the switching function, in case of the first order phase transition, goes to 0 as N goes to infinity. In a first-order phase transition the transition part of the function becomes more and more steep as the number of microscopic units (atoms, spins, monomers) grows, and finally becomes a step function. Indeed, as $N? \to \infty$, the free energy difference between the phases (local minima of the landscape) varies as $\Delta F \sim N\,(T - T^{\star})$. This is because phase transition is a cooperative phenomenon, it is driven by the interactions between particles. Thus, the free energy difference, as a function of the temperature, roughly speaking, is either zero or infinite (in thermodynamic limit), that is the probability to be in a phase switches from zero to one. That means, the sigmoidal temperature dependence is degenerated into a step function. In chemical systems, there is no cooperativity between the microscopic units, and therefore the switching is soft, it is not a step but a sigmoidal curve.

Proteins, and protein folding, belong somehow to both worlds of chemical reactions and of first order transitions. Protein folding is similar to a chemical reaction because folding is the event that involves just one protein molecule, and there is no cooperativity between protein molecules. But, protein folding is also close to a first order transition, because protein molecule contains a large number of monomers N, and there is strong cooperativity between monomers in the protein molecule.

Kinetics of protein folding includes, first of all, polymer chain collapse and also self-organization of the "correct" $3D$ structure. It is known that some proteins first collapse and then search for the native state through the re-arrangements of the collapsed globule, while others wait longer in the open coil state, but then jump into a more or less correct globular state. Initial stages of folding, initial collapse of the coil, seems to be insensitive to the details of the heteropolymer sequence. This argument suggests that these stages should be well modeled by a homopolymer collapse.

3.7 Reaction Coordinate and Levinthal Paradox

Our discussion of the chemical reactions and first order phase transitions allows us to look at the Levinthal paradox from a new refreshing point of view.

As a starting point, it is convenient to imagine a plot showing free energy as a function of an order parameter, or reaction coordinate, x. As we discussed, it must have two minima separated by a maximum. When we are close to the first order transition, the dependence $F(x)$ evolves with changing temperature such that one minimum gets deeper, while the other one gets more shallow. At the transition point, they switch the roles of the deepest and the less deep ones. This free energy can be defined if one keeps the value of an order parameter, x, fixed, while allowing all other degrees of freedom to relax to thermodynamic equilibrium. For equilibrium statistical mechanics, this consideration is OK, and basically anything can serve as an order parameter, provided that it has different values in two different phases. In kinetics, the situation is far more delicate.

Indeed, let us consider temperature at which both minima of free energy $F(x)$ are at roughly the same depth. They are separated by a barrier. It looks like the system has to overcome THIS barrier to go from one phase to another. Then, the transition time

should be proportional to $\tau \exp(\Delta F/k_B T)$, where $\Delta F = \max[F(x)] - \min[F(x)]$ is the barrier height. However, as free energy is an extensive quantity, the height of the barrier, ΔF, is proportional to N and goes to infinity in thermodynamic limit. In fact, this is just another formulation of Levinthal paradox, because we arrived at the transition time which is exponential in N.

Luckily, there is nothing specific for proteins in this consideration, and we can turn for lessons to other, better understood fields of physics. Consider, for instance, gas-liquid phase transition, perhaps the simplest first order phase transition. For the equilibrium theory of this transition, the most natural order parameter is the overall averaged density of the system. As a function of density, free energy has exactly the shape that we discussed, with two minima separated by a maximum. What is the nature of the maximum? If we start from a gas and increase its density, say, to half that of liquid, then we are at disadvantage from both energy and entropy view points: energetically, our molecules are not close enough to enjoy attractive interactions, so we gained nothing; entropically, density is much higher than in the gas, so we lost a lot. Thus the maximum.

So, how does the gas condenses? This occurs through nucleation. That means, the system does not progress through a succession of overall uniform states with gradually increasing density, instead it goes through the states with dramatically non-uniform density, which is the droplets of liquid in the gas. In this sense, density is not an appropriate order parameter, or not a good reaction coordinate in chemical kinetics language.

What turns out a better choice is the radius of the nucleus. We imagine that a nucleus is a spherical shaped droplet of a new phase, and look at its radius, R. As a function of R, free energy is usually written in the form $\Delta F(R) = -\alpha(4/3)\pi R^3 + \gamma 4\pi R^2$, where the first term is gain in free energy due to the formation of the new phase, its proportional to the volume of the nucleus, while the latter term is the surface energy of the contact between the droplet and the surrounding gas, it is unfavorable (positive), and proportional to the surface area. It is easy to see, that as a function of R this free energy, $\Delta F(R)$, has a characteristic shape with the maximum. It is this maximum, the theory says, that has to be overcome. Obviously, it has nothing to do with N, it does not diverge in thermodynamic limit. It corresponds to the so-called critical nucleus size. In the beginning, the system has to climb on the barrier, it must be a thermo-activated process, entirely due to fluctuations. However, as soon as the nucleus is larger than critical, it starts sliding down the potential hill.

What does this theory assume? Although this is rarely emphasized, this theory assumes that the radius R is a very slow variable. Indeed, this is the only assumption that allows to write down $\Delta F(R)$ as we did before. For instance, by using the surface tension we assume that the surface structure achieves thermodynamic equilibrium at every value of R. By the same token, we assume that the shape of nucleus relaxes much faster than its size, etc. By the way, recent simulations by Daan Frenkel and his co-workers[19] suggest that all these assumptions are not so benign and not universally applicable even for the gas-liquid transition.

Returning to the general consideration, the problem is easy if there is a variable, let say x, which relaxes much slower than all other variables. Then, at the fixed x, the system has time to equilibrate all of its other degrees of freedom, and in this case our considerations become applicable kinetically. That is the problem of the choice of reaction coordinate, or kinetic order parameter. It is a difficult problem, which does not have known general solution.

3.8 Commitment as the Best Possible Reaction Coordinate

Let us assume that we can calculate for each point in the conformation space the probability that the protein is going to fold before ever touching the unfolded state. This way, we can formally define a function $P_{\rm fold}$ over the conformation space[20–22]. This probability is almost unity around the native state, and is almost zero around the unfolded states. This probability is also called commitment, it measures, for every conformation, to which extent this conformation is committed to fold.

Let us first discuss this quantity for a lattice model[20]. To understand what is $p_{\rm fold}$, we assume that there are two well defined states of the system, in case of lattice toy proteins those are folded and unfolded states. For the present discussion, it is not important that folded state is usually represented by just a single compact conformation, while unfolded state is a big ensemble of coil-like conformations. What is important is that in the conformation space graph certain vortices are labeled as belonging to unfolded state, while certain others are labeled as belonging to the folded state.

Now, we pick an arbitrary conformation, or an arbitrary vertex of the conformation space graph, and define $p_{\rm fold}$ for this conformation in the following way. Imagine that we perform many Monte Carlo runs using the chosen conformation as an initial one. Every time, we run Monte Carlo dynamics for as long as it takes for the system to arrive for the first time into either folded or unfolded state. Then, we have to collect the statistics over many runs, and determine the probability, p, that a run, or randomly chosen trajectory, will first arrive into the folded state before ever touching the unfolded state. This probability is $p_{\rm fold}$.

Quite similarly, commitment can be defined for an off-lattice system[21,22]. In this case, we choose a particular configuration (or conformation), and then initiate many trajectories going out of this point by randomly choosing the full set of momenta (or velocities). Once again, we determine the probability that the trajectory first hits the folded state before ever touching the unfolded state.

What is the meaning of this quantity, whether one prefers calling it $p_{\rm fold}$ or commitment? If p is large, close to unity, that means the system is very close to the folded state, its descend back to the unfolded state is unlikely. Vice versa, if p is small, that means the system is close to the unfolded state.

We shall argue that $p_{\rm fold}$, or commitment, is, in principle, THE BEST possible REACTION COORDINATE, or an order parameter. Of course, practical disadvantage of the value $p_{\rm fold}$ is that it is difficult to compute and it has not obvious meaning, it is not anything like nucleus radius. But as a matter of principle, there cannot be better reaction coordinate. We shall prove this fact considering kinetics of the lattice model. This is essentially the random walk of a particle on the graph.

Indeed, if there were the slowest degree of freedom in the system, then p would be clearly a well defined monotonic function of this coordinate. As the system progresses along the selected single degree of freedom from an unfolded to the folded state, its value of p necessarily increases. Thus, if the reaction coordinate in the traditional sense existed, it would be basically the same as commitment. However, p can be computed by a constructive algorithm, albeit time consuming one, while reaction coordinate is never known. And, most importantly, $p_{\rm fold}$, or commitment, remains well defined even for the generic system, with no special slow degree of freedom.

The concept of commitment, or $p_{\rm fold}$ is currently widely used in protein folding simulations (see, for instance,[23]). Nevertheless, there is a feeling that better intuition about this quantity is necessary. For instance, the works on the statistical mechanics of folding trajectories[24] lead to the questions like the ones about local extrema of $p_{\rm fold}$. Is it possible to find a state such that all surrounding states have $p_{\rm fold}$ smaller (or larger) than the given one? If the answer were to be positive, that would render the concept of commitment useless at the very least. Luckily, the intuition suggests that it should be impossible. Below, we shall prove that it is impossible indeed.

3.9 Direct Current Analogy

In this section, we shall describe the physical analogy which helps intuitive understanding of the commitment, and also allows to prove their general properties, like monotonous behavior mentioned in the previous paragraph. This analogy is presented in terms of direct currents governed by the Kirchoff rules. To formulate it, let us be a little more specific about the model under study.

We shall consider Metropolis Monte Carlo dynamics - the method first described in a 1953 paper[25] by Nicholas Metropolis, Arianna Rosenbluth, Marshall Rosenbluth, Augusta Teller, and Edward Teller. The idea of the method was reportedly conceived by Metropolis, Rosenbluth, and Teller during a Los Alamos dinner party.

Suppose our polymer has certain available conformations, and let us call the conformations with letter $\mathcal{C}$. There is potential energy landscape, which means that there is certain energy $U(\mathcal{C})$ associated with every conformation $\mathcal{C}$. Consider now a pair of conformations, say $\mathcal{C}$ and $\mathcal{C}'$, such that they are connected on the conformation graph; in other words, they can be transformed one into another via a single Monte Carlo move. Consider the flow between $\mathcal{C}$ and $\mathcal{C}'$. Of course, this means, we should imagine a large number (an ensemble) of computers independently running Monte Carlo simulation of our system, and then we should ask what is number of copies switching from $\mathcal{C}$ to $\mathcal{C}'$ at the given time? Equivalently, we can speak of a probability to find a single Monte Carlo process jumping from $\mathcal{C}$ to $\mathcal{C}'$ at the given time. To be specific, let us assume that $\mathcal{C}$ is higher in energy than $\mathcal{C}'$: $U(\mathcal{C}) > U(\mathcal{C}')$. Then, Metropolis criteria[25] accepts the move from $\mathcal{C}$ to $\mathcal{C}'$ every time the move is offered, we get that the flow from $\mathcal{C}$ to $\mathcal{C}'$ is simply proportional to the occupation number of $\mathcal{C}$ at the present time, $n_{\mathcal{C}}$. As regards the opposite moves from $\mathcal{C}'$ to $\mathcal{C}$, they are accepted with probability smaller than unity, namely $\exp\left[U(\mathcal{C}') - U(\mathcal{C})\right]$. We should also keep in mind that the connectivity of the conformation graph in different points might be different. Specifically, let us say that every conformation $\mathcal{C}$ has some $\mu_{\mathcal{C}}$ neighbors on the graph. Then, the bond leading from $\mathcal{C}$ to $\mathcal{C}'$ is offered with the probability $1/\mu_{\mathcal{C}}$. Therefore, we can summarize all this by writing the following relation for the flux of probability from $\mathcal{C}$ to $\mathcal{C}'$:

$$I_{\mathcal{C}\to\mathcal{C}'} = \frac{n_{\mathcal{C}}}{\mu_{\mathcal{C}}} \min\left\{1; \frac{\mu_{\mathcal{C}}}{\mu_{\mathcal{C}'}} \exp\left[\frac{U\left(\mathcal{C}\right) - U\left(\mathcal{C}'\right)}{k_B T}\right]\right\} . \tag{1}$$

Similarly, the flux in the opposite direction is

$$I_{\mathcal{C}'\to\mathcal{C}} = \frac{n_{\mathcal{C}'}}{\mu_{\mathcal{C}'}} \min\left\{1; \frac{\mu_{\mathcal{C}'}}{\mu_{\mathcal{C}}} \exp\left[\frac{U\left(\mathcal{C}'\right) - U\left(\mathcal{C}\right)}{k_B T}\right]\right\} . \tag{2}$$

Below, to save some writing, we shall omit the temperature factor, assuming $k_BT = 1$; in other words, we pretend that potential energy $U(\mathcal{C})$ is measured in the units of k_BT: $U(\mathcal{C}) \to U(\mathcal{C})/k_BT$.

It is convenient to rewrite equations (1) and (2) by introducing the following notations. For every conformation $\mathcal{C}$, we define the quantity

$$\phi_{\mathcal{C}} = n_{\mathcal{C}} e^{U(\mathcal{C})} \ , \tag{3}$$

and for every connection between $\mathcal{C}$ and $\mathcal{C}'$, we define another quantity

$$R_{\mathcal{C}\mathcal{C}'} = \max\left\{\mu_{\mathcal{C}} e^{U(\mathcal{C})}; \mu_{\mathcal{C}'} e^{U(\mathcal{C}')}\right\} \ . \tag{4}$$

One nice thing about quantities $\phi_{\mathcal{C}}$ is that in equilibrium, when occupation numbers $n_{\mathcal{C}}$ are governed by Boltzmann distribution $n_{\mathcal{C}} \propto e^{-U(\mathcal{C})}$, the values of $\phi_{\mathcal{C}}$ become independent of $\mathcal{C}$, a constant all over the conformation space. Using $\phi_{\mathcal{C}}$ and $R_{\mathcal{C}\mathcal{C}'}$, the master equations (1) and (2) can be presented in the form

$$I_{\mathcal{C}\mathcal{C}'} \equiv I_{\mathcal{C}\to\mathcal{C}'} - I_{\mathcal{C}'\to\mathcal{C}} = \frac{\phi_{\mathcal{C}} - \phi_{\mathcal{C}'}}{R_{\mathcal{C}\mathcal{C}'}} \ . \tag{5}$$

In this formula, it is now easy to recognize the Ohm's law for direct currents, which leads to the following physical interpretation.

We imagine that the conformational graph of the system is a network of resistors. The resistance between nodes $\mathcal{C}$ and $\mathcal{C}'$ is $R_{\mathcal{C}\mathcal{C}'}$. Governed by the Kirchoff rules, there are potentials on each node, $\phi_{\mathcal{C}}$, and currents between the nodes, $I_{\mathcal{C}\mathcal{C}'}$. The knowledge of these potentials and the current distribution is equivalent to the full knowledge of the dynamics of the system in question. Thus, we see that occupation numbers (or probabilities) $n_{\mathcal{C}}(t)$ satisfy the equation which is an analog of diffusion equation:

$$\frac{\partial n_{\mathcal{C}}(t)}{\partial t} = -\sum_{\mathcal{C}' \text{ at } \mathcal{C}} I_{\mathcal{C}\mathcal{C}'} \ , \tag{6}$$

where the summation runs over all sites $\mathcal{C}'$ which are neighbors of the site $\mathcal{C}$ on the graph.

In fact, as we know from the experience with the resistor networks, the important ingredient of the problem is the externally applied voltage. In the presence of such applied voltage, direct currents may flow and remain time-independent. The easiest way here is to look at the analog of Green's function of the diffusion equation, which is equivalent to applying a δ-type voltage source at just one point, say, $\mathcal{C}_0$:

$$\frac{\partial n_{\mathcal{C}}(t)}{\partial t} = -\sum_{\mathcal{C}' \text{ at } \mathcal{C}} I_{\mathcal{C}\mathcal{C}'} + Q\delta(\mathcal{C}_0)\delta(t) \ . \tag{7}$$

It is useful now to realize the more specific meaning of occupation numbers $n_{\mathcal{C}}(t)$ in our specific context. One way to understand them is to imagine that there are many processors simultaneously running the Monte Carlo simulation of a given protein. Then, $n_{\mathcal{C}}(t)$ is the fraction of all processors whose simulations are at the point $\mathcal{C}$ at the time moment t. From that point of view, we should imagine the rule best adjusted to computing the $p_{\rm fold}$ or commitment: every time that a particular computer achieves folding, it is re-launched again starting from the point $\mathcal{C}_0$. In this setting, we will eventually achieve a stationary distribution with time-independent $n_{\mathcal{C}}$, but with non-zero and non-trivial currents. The equation

for such stationary distribution is derived by time integration of the diffusion equation (7):

$$0 = -\sum_{\mathcal{C}' \text{ at } \mathcal{C}} I_{\mathcal{C}\mathcal{C}'} + Q\delta(\mathcal{C}_0) \,, \tag{8}$$

or, in the other words,

$$\sum_{\mathcal{C}' \text{ at } \mathcal{C}} \frac{n_{\mathcal{C}} e^{U(\mathcal{C})} - n_{\mathcal{C}'} e^{U(\mathcal{C}')}}{\max\left\{\mu_{\mathcal{C}} e^{U(\mathcal{C})}; \mu_{\mathcal{C}'} e^{U(\mathcal{C}')}\right\}} = Q\delta(\mathcal{C}_0) \,, \tag{9}$$

3.10 Direct Current Formulation of the First Return Problem and its Easy Solution

The concept of commitment, as it is formulated above, is based on the classical probabilistic "first return" problem. Indeed, commitment of the state $\mathcal{C}$ is the probability that the trajectory of the random walk on the graph, starting from $\mathcal{C}$ will arrive *for the first time* into the folded state $\mathcal{F}$ before ever hitting the unfolded state $\mathcal{U}$. Let us consider the simple classical example[26].

Suppose a gambler arrives at the casino with certain amount of money x_0 and keeps gambling indefinitely unless arriving at the desperate zero money state ($x = 0$), in which case he/she is discarded. The question is this: what is the probability that the player looses money (arrives for the first time at $x = 0$) at the time t? The standard way to solve this problem is to realize that the time-dependent probability distribution of the money at possession of the gambler satisfies the diffusion equation:

$$\frac{\partial n_x(t)}{\partial t} = D\frac{\partial^2 n_x(t)}{\partial x^2} + \delta(t)\delta(x - x_0) \,, \tag{10}$$

where D is an appropriate diffusion coefficient (which can be absorbed into the time measurement units), and δ-functions describe the initial condition. The central idea of this approach is to realize that the concept of *first return* is embodied in the absorbing boundary condition:

$$n_x(t)|_{x=0} = 0 \,. \tag{11}$$

This boundary condition ensures that in the path integral solution of the diffusion equation, all trajectories are discarded which ever visit the left half-line ($x < 0$). Now, assuming $n_x(t)$ is found, the probability to arrive at $x = 0$ at time t *for the first time* is given by the flux into $x = 0$: $W(t) = -D\partial n_x(t)/\partial x|_{x=0}$. The corresponding solution is not difficult to find:

$$W(t) = \frac{x_0}{2\sqrt{\pi D}t^{3/2}} e^{-x_0^2/4Dt} \,, \tag{12}$$

and the classical theorem[26] says that $\int_0^\infty W(t)dt = 1$: sooner or later, gambler looses all money with probability one.

Here is now the simple solution of this problem based on the direct current analogy (see also[27]). Consider semi-infinite line of resistors (or just a wire of uniform resistance) going along the x-axis. In terms of analogy, the absorbing boundary condition (11) means that the boundary is grounded, i.e., is kept at the zero potential (3). Suppose now we feed a direct current into the point x_0 (of course, applying a constant voltage to the point x_0). Then, the whole statement of the theorem is that the current leaving the system through the

grounded terminal at $x = 0$ is equal to the current entering the system through the terminal at x_0 - which is trivial.

Of course, this simple solution comes from the use of stationary equation (9.

3.11 Direct Current Formulation of the Commitment

Now we are prepared to formulate the way to determine the commitment of any given state $\mathcal{C}$ using the direct current analogy. We understand that the first arrival condition can be imposed by absorbing boundary conditions, or grounding the corresponding sites. Therefore, the formulation is as follows.

Suppose all the sites $\mathcal{U}$ corresponding to the unfolded state are grounded, as well as the site $\mathcal{F}$ corresponding to the folded state. Suppose further that we feed a direct (stationary, time independent) current I into the site $\mathcal{C}$. This current flows partly to $\mathcal{U}$ and partly to $\mathcal{F}$; in obvious notations, $I = I_{\mathcal{U}} + I_{\mathcal{F}}$. Then, the commitment is nothing but the fraction of current going into the $\mathcal{F}$ terminal. According to the Ohm's law, this can also be written in terms of the corresponding resistances:

$$p_{\rm fold}(\mathcal{C}) = \frac{I_{\mathcal{F}}}{I} = \frac{R_{\mathcal{C}\mathcal{U}}}{R_{\mathcal{C}\mathcal{U}} + R_{\mathcal{C}\mathcal{F}}} \, . \tag{13}$$

It is instructive to consider here a simple example, which is the diffusion in one-dimensional potential landscape $U(x)$. Of course, in this case the choice of reaction coordinate is trivial, there nothing but x. Nevertheless, we can imagine that there are two ends of the diffusion interval, which we call points $\mathcal{U}$ and $\mathcal{F}$, and we want to know what is the probability to start from some point $\mathcal{C}$ between $\mathcal{U}$ and $\mathcal{F}$ and arrive at $\mathcal{F}$ for the first time without ever touching $\mathcal{U}$. In this case, the resistor network is one-dimensional, all resistors are connected in series. Remembering the expression (4), we see momentarily that formula (13) yields

$$p_{\rm fold}(\mathcal{C}) = \frac{\int_{\mathcal{C}}^{\mathcal{U}} e^{U(x)} dx}{\int_{\mathcal{F}}^{\mathcal{U}} e^{U(x)} dx} \, . \tag{14}$$

It is clearly seen both in the example (14) and in the general formula (13) that the commitment is always positive (non-negative) and never exceeds unity - as it must be for the probability.

3.12 Direct Current Formulation of the Landscape

Our result (13) generates an insight into the general meaning of commitment and its properties. For instance, we can proceed in the following way. Let us ask what are the surfaces of $p_{\rm fold} = {\rm const}$, and what is their topology.

To understand it, let us imagine yet another experiment with the direct current. Let us ground this time only the folded state site $\mathcal{F}$, and let us feed the current into the unfolded state sites $\mathcal{U}$. In other words, the potentials of the terminals at the folded and unfolded states are equal to $\phi_{\mathcal{F}} = 0$ and $\phi_{\mathcal{U}} = V$, respectively. In this case, according to the Ohm's law, the potential at the arbitrary site $\mathcal{C}$ should be equal to

$$\frac{\phi_{\mathcal{C}}}{V} = \frac{R_{\mathcal{C}\mathcal{F}}}{R_{\mathcal{C}\mathcal{U}} + R_{\mathcal{C}\mathcal{F}}} = 1 - p_{\rm fold} \, . \tag{15}$$

Thus, this potential essentially gives us the value of the commitment $p_{\rm fold}$. Of course, this realization of the direct current model corresponds very directly to the most common computer experiment in which we start from somewhere in the unfolded region and look for the first arrival into the folding state.

The result (15) indicates that the topology of every surface of constant commitment is such that it separates folded state from unfolded state; moreover, and more general, every surface $p_{\rm fold} = {\rm const} = p_0$ separates the regions with $p < p_0$ and with $p > p_0$. In particular, the commitment has no local maximum or local minimum. Its only minimum is the global one at the unfolded state, where it is equal to zero, and its only maximum is that at the folded state where it is equal to unity. This confirms the intuition according to which the commitment can be used as a reaction coordinate.

Now, the central question arises: what is the relevant free energy profile if we choose to use $p_{\rm fold}$ as the reaction coordinate? It turns out possible to address this question in quite general form. In this argument, we shall disregard the discrete character of conformation space graph and pretend to work with the continuous conformations space. Moreover, we shall assume that the conformation space is flat, which means that $\mu_{\mathcal{C}}$ does not depend on $\mathcal{C}$.

Consider our latest direct current formulation, in which terminals are at the folded and the unfolded states, and equipotential surfaces are simultaneously the surfaces of constant commitment. Since the current does not flow along the equipotential surface, we can ask what is the resistance of the conformation space layer sitting between the surfaces $p_{\rm fold} = p$ and $p_{\rm fold} = p + dp$. Since these surfaces are equipotential, all resistors connecting them may be viewed as connected in parallel. For them, the conductivities must be summed together, which yields

$$\frac{1}{R(p)} = \sum_{\mathcal{C}, p={\rm const}} \frac{1}{R_{\mathcal{C}\mathcal{C}'}} \,. \tag{16}$$

Remembering the formula (4), it is natural to define $U_{\rm eff}(p)$ according to $R(p) = e^{U_{\rm eff}(p)}$, and then we obtain

$$U_{\rm eff}(p) = -\ln \left[\sum_{\mathcal{C}, p={\rm const}} e^{-U(\mathcal{C})} \right] \,. \tag{17}$$

Strikingly, this formula coincides with the definition of the partially equilibrium free energy as a function of order parameter, except it does not involve any assumptions regarding slow degree of freedom and the like. Instead of being the definition of the statistical mechanics partition function, formula (17) describes the parallel connection of the resistors. What this formula proves is that one can use the commitment as the reaction coordinate, with the relevant free energy profile given by the formula (17). In particular, the concept of the folding barrier is clearly formulated as the place with highest resistance.

It should be born in mind that equation (17) deserves further attention. In particular, it remains unclear how we should proceed with generalizing it for the case when $\mu_{\mathcal{C}}$ does depend on $\mathcal{C}$. This raises the question of the geometry of the real conformation space.

4 Compact Conformations: Lattice Model

The study of the geometry of compact polymers is quite tricky. The theory of random compact conformations is well developed on the mean field level (see, e.g., in the book[28]). This is the theory of homopolymer globules, because they are entropically dominated by the most typical conformations. Major conclusion of the mean field theory is that chain segments inside the globule follow Gaussian statistics, and do not exhibit any signs of order. This conclusion is in sharp contradiction with the statements in the literature[29–31] that compactness of the conformation may favor elements of secondary structures, such as α-helices and β-pins.

Computationally, the problem of compact conformations is closely related to that of Hamiltonian walks on the graphs. We remind the reader that the concept of a Hamiltonian walk was introduced by Hamilton in connection with famous Euler problem of Königsberg bridges: the task was to find the Sunday promenade passing every one of the seven bridges, never returning to the already visited place. In general, Hamiltonian walk on an arbitrary graph can be defined as a walk which visits every site on the graph once and only once. If our graph is, say, $\ell \times m \times n$ piece of the cubic lattice in $3D$, then Hamiltonian walk on such graph is the same as maximally compact conformation of the polymer filling $\ell \times m \times n$ domain.

Enumeration of Hamiltonian walks on graphs is well known problem in combinatorics. Of course, the best possible statistics is achieved by exhaustive enumeration of all Hamiltonian walks. This is possible for rather short polymer chains only: for the chains with 27 monomers filling $3 \times 3 \times 3$ of the cubic lattice[32], and also for 36- and 48-mers, filling $3 \times 3 \times 4$ and $3 \times 4 \times 4$ segments, respectively[33]. Obviously, these chains are far too short to address statistics and fractal structure of the typical conformation.

Short of exhaustive enumeration, other methods to generate larger compact conformations have been suggested. The most straightforward Monte Carlo chain growth methods[34] are totally inefficient for long compact chains, because of catastrophic explosion of rejected looped conformations. Transfer matrix approach put forward by[35–37] is very efficient for the chains filling an elongated domain $\ell \times m \times n$, where one of the dimensions, say n, may be arbitrarily large. Unfortunately, to remain within computational tractability, two other dimensions, ℓ and m, must be small, not greater than 2 or 3. An alternative approach, suggested in[38], is free of this limitation. It employs combinatorial techniques of two-matching and patching of bipartite graphs. Unfortunately, we found that this method generates conformations in a heavily biased way.

In the forthcoming paper[39], we demonstrate significant improvements to the algorithm of Ramakrishnan *et al*[38] Compact conformations (Hamiltonian walks) and cycles were generated up to the size $22 \times 22 \times 22$. Their fractal and topological properties were examined in details. The major findings were as follows.

First, it was confirmed that local fractal structure of typical compact conformations is Gaussian at the scale smaller than the globule size. In other words, it satisfies Flory theorem. Second, it was found that for linear compact chain the correlation between chain ends is minimal, it is basically reduced to the excluded volume effect (both ends cannot occupy the same place). Third, it was found that chain compaction quite significantly increases the probability of forming non-trivial knots. These results should be born in mind when we consider conformations of real proteins and try to analyze whether they are random or not.

4.1 Native Conformations

So far we though about a native state of a protein in quite abstract way. In fact, several thousands real native conformations are now known due to the X-ray diffraction analysis. Surprisingly, many of the native conformations are quite alike despite they belong to totally different proteins. This might be the result of either convergent or divergent evolution. In the former case, we imagine that certain conformations have some evolutionary advantages, and then evolutionary development bring proteins to have these advantageous conformations. Another possibility is to imagine that similarities between conformations of proteins might be due to the common ancestry of these proteins. Not going too deeply into this exciting subject, we only make a reference to the interesting recent paper[40] and the references therein.

5 Conclusion

It should be noted that the understanding of Levinthal paradox has progressed very significantly since it was first formulated[14]. First of all, it is found that the folding time, under the conditions of thermodynamic equilibrium between folded and unfolded states, scales as $\tau \exp\left(s' N^{2/3}\right)$[41,42], which is very significantly smaller than Levinthal time proportional to $\tau \exp\left(sN\right)$. This estimate, as already said, is valid under the conditions of thermodynamic equilibrium, which means that it relies on the transition between denatured and native forms being highly cooperative, of all-or-none type. Indeed, high cooperativity is a well established experimental fact[2]. It is also well understood that high cooperativity is the property of proteins which is due to their peculiar selected sequences. Among random sequences, vast majority would not have exhibited any signs of cooperativity, as it was first established by Shakhnovich and Gutin[43]. This latter fact has been extensively tested using lattice models (as described, e.g., in the review article[44]; see also references therein). Since everything related to the lattice models is perceived with a large dose of (healthy?) skepticism in protein community, it is important to emphasize that the fact of non-cooperative folding in the majority of sequences is well understood beyond lattice models. Actually, it was foreseen by Bryngelson and Wolynes a long time ago[45].

Speaking about the relation between sequence selection, the all-or-none cooperative mechanism of folding, as the scaling of folding time, it is interesting to mention that experimental observations do not provide any evidence on the folding (under equilibrium conditions) time dependence on the chain length. That means, the above mentioned theoretical prediction, $\tau \exp\left(s' N^{2/3}\right)$, although sufficient to rule out any paradoxes, maybe still an overestimate.

The role of sequence selection is also well understood from a different view point, namely, related to the mutation stability (see also in the review article[44]). In the majority of sequences, every mutation breaks the stability of the native state with the probability very close to 100%. By contrast, the selected sequences - the same ones which exhibit highly cooperative folding-unfolding transition! - are reliable in the sense that their native state with high probability survives and remains stable even after several mutations.

While the real mechanisms of evolutionary sequence selection remain unknown, and while the computational models of sequence selection keep improving since the first suggestions[46,47], it is getting increasingly clear that there are many sequences which meet the sufficient criteria of reliable folding.

To summarize, in the light of all the findings of the last decade, it seems clear (to the present author, at least) that the discussions about Levinthal paradox must now move forward to the new spheres. How does the sequence selection work (or worked) in real evolution? What are the specific scenario of folding dynamics for selected sequences - how specific is the nucleation, how many and which conformations belong to the transition state, what is the reaction coordinate associated with folding; in other words, how precisely do these selected sequences slide down their folding funnels[48,23,49,50]? What are the physical principles behind the selection of certain spatial structures, or folds and fold families[51]? What are the general physical principles behind the enzymatic, motor and other functions of proteins, and do they have any relation to the principles involved in folding? What are the mechanisms of aggregation, or mechanisms preventing aggregation, of proteins? There are very many works on these subjects, to make the list of them is a daunting task far beyond the framework of the present note. However many questions remain open, it seems that the Levinthal's question - how can protein sample "biologically large" number of conformations - has been answered: protein does not sample them. Most importantly, there are sufficiently many "good" sequences for the evolution to select from, where every "good" sequence is capable of folding, and does not need exhaustive conformation sampling to do so. Understanding this was a remarkable achievement of the last decade.

References

1. A. V. Finkelstein, O. B. Ptitsyn *Protein Physics*, Academic Press, 2002.
2. P.Privalov Adv. Protein. Chem. **33**, 167-241 (1997).
3. C.Anfinsen, R.Redfield, W.Choate, J.Page, W.Carroll *Studies on the Gross Structure, Cross-Linkages, and Terminal Sequences in Ribonuclease* Journal of Biological Chemistry **207**, 201-210 (1954).
4. Christian B. Anfinsen *Studies on the Principles That Govern the Folding of Protein Chains, The Nobel Lecture delivered on 11 December 1972*. Les Prix Nobel en 1972, p. 103-119, Nobel Foundation, 1973.
5. H. Fraunfelder, Colloquium talk at the University of Minnesota Physics Department, October, 2002.
6. O. B.Ptitsyn, M. V.Volkenstein Journal of Biomolecular Structure and Dynamics **4**, 137 (1986).
7. P. H. Verdier, W. H. Stockmmayer Journal Chem. Phys. **36**, 227 (1962).
8. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, P. G. Wolynes: Proteins **21**, 167-195 (1995).
9. M. R. Betancourt, J. N. Onuchic Journal Chem. Phys. **103**, 773-787 (1995).
10. R. Du, A. Grosberg, T. Tanaka *Random Walks in the Space of Conformations of Toy Proteins* Phys. Rev. Letters **84**, 1828-1831 (2000).
11. A. Scala, L. A. Nunes Amaral, M Barthèlèmy Europhys. Lett. **55**, 594-600 (2001).
12. R. Du, A. Grosberg, T. Tanaka, M. Rubinstein *Unexpected Scenario of Glass Transition in Polymer Globules: an Exactly Enumerable Model* Phys. Rev. Letters **84**, 2417-2420 (2000).
13. E. Shnol, preprint, Pouschino (1980).
14. C.Levinthal Journal Chim. Phys. Biol. **65**, 44-45 (1968).
15. I.Berezovsky, E.Trifonov *Loop fold structure of proteins: resolution of Levinthal's*

paradox J. Biomol. Struct. Dyn. **20**, 5-6 (2002); A.Finkelstein *Cunning simplicity of hierarchial folding* Journal Biomol. Struct. Dyn. **20**, 311-313 (2002); I.Berezovsky, E.Trifonov *Back to units of protein folding* Journal Biomol. Struct. Dyn. **20**, 315-316 (2002); A.Grosberg *A few disconnected notes related to Levinthal paradox* Journal Biomol. Struct. Dyn. **20**, 317-321 (2002); A.Kloczkowski, R.Jernigan *Loop folds and evolutionary conservation of folding nuclei* Journal Biomol. Struct. Dyn. **20**, 323-325 (2002); M.Rooman, Y.Dehouck, J.Kwasigroch, C.Biot *What is paradoxical about Levinthal paradox?* Journal Biomol. Struct. Dyn. **20**, 327-329 (2002); A.Fernandez, A.Belinky, M. de las Mercedes Boland *Protein folding: where is the paradox?* Journal Biomol. Struct. Dyn. **20**, 331 (2002).

16. E. Lifshitz, L. Pitaevskii: *Physical Kinetics*, reprint edition (Butterworth-Heinemann, 1981).
17. T. C. B.McLeash *Protein folding in high-dimensional spaces: hypergutters and the role of non-native interactions* `http://arxiv.org/abs/physics/0311028`.
18. P. Hanggi, P. Talkner, M. Borkovec *Reaction-rate theory: fifty years after Kramers* Reviews of Modern Physics **62**, 251 (1990).
19. S. Auer, D. Frenkel *Prediction of absolute crystal-nucleation rate in hard sphere colloids* Nature **409**, 1020-1023 (2001).
20. R. Du, V. Pande, A. Grosberg, T. Tanaka, E. Shakhnovich *On the Transition Coordinate for Protein Folding* Journal Chem. Phys. **108**, 334-350 (1998).
21. P. Bolhuis, C. Dellago, P. Geissler, D. Chandler: *Transition path sampling: throwing ropes over mountains in the dark* Journal Phys. - Condensed Matter **12**, A147-A152 (2000).
22. P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, *Transition path sampling: Throwing ropes over rough mountain passes, in the dark* Annu. Rev. Phys. Chem. **53**, 291-318 (2002).
23. V. Pande, A. Grosberg, T. Tanaka, D. Rokhsar *Pathways for Protein Folding: is a New View Needed?* Current Opinion in Structural Biology **8**, 68-79 (1998).
24. P. Eastman, N. Grönbech-Jensen, S. Doniach *Simulation of protein foldeing by reaction path annealing* Journal Chem. Phys. **114**, 3823 (2001).
25. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller *Equation of state calculations by fast computing machines* Journal Phys. Chem. **21**, 1087-1092 (1953).
26. W. Feller: *An introduction to probability theory and its applications*, 2nd edn. (J. Wiley & Sons, New York 1971).
27. S. Redner: *A guide to first passage processes* (Cambridge Univ. Press, 2001).
28. A. Yu. Grosberg, A. R. Khokhlov, *Statistical Physics of Macromolecules*, AIP Press, New York, (1994).
29. D.Yee, H.Chan, T.Havel, K.Dill Journal Mol. Biol. **241**, 557-573 (1994).
30. A. Maritan, C. Micheletti, A. Trovato, J. R. Banavar, Nature **406(6793)**, 287-290 (2000).
31. J. Banavar, A. Maritan *Colloquium: Geometrical approach to protein folding: a tube picture* Reviews of Modern Physics **75**, 23-34 (2003).
32. E. Shakhnovich, A. Gutin Journal Chem. Phys. **93**, 5967 (1990).
33. V. S. Pande, C. Joerg, A. Yu. Grosberg, T. Tanaka, Journal Phys. A: Math. Gen. **27(18)**, 6231-6236 (1994).

34. *Markov Chains and Monte Carlo Calculations in Polymer Science, Monographs in Macromolecular Chemistry*, edited by G. G. Lowry, Marcel-Dekker, New York, (1970).
35. A. Kloczkowski, R. L. Jernigan Journal Chem. Phys. **109(12)**, 5134-5146 (1998).
36. A. Kloczkowski, R. L. Jernigan Journal Chem. Phys. **109(12)**, 5147-5159 (1998).
37. A. Kloczkowski, R. L. Jernigan Macromolecules **30(21)**, 6691-6694 (1997).
38. R. Ramakrishnan, J. F. Pekny, J. M. Caruthers Journal Chem. Phys. **103(17)**, 7592 (1995).
39. R.Lua, A.Borovinskiy, A.Grosberg *Fractal and statistical properties of large compact polymers: a computational study* Polymer **45(2)**, 717-731 (2004).
40. N. V. Dokholyan, B. E. Shakhnovich, E. I. Shakhnovich *Expanding protein universe and its origin from the biological Big Bang* Proc. Nat. Acad. Sci. USA **99**, 14132-14136 (2002).
41. A.Finkelstein, A.Badretdinov Folding & Design **2**, 115 (1997); Folding & Design **3**, 67 (1998).
42. R.Du, A.Grosberg, T.Tanaka *Coexistence of Native and Denatured Phases in a Single Protein-like Molecule* Phys. Rev. Letters **83**, 4670-4673 (1999).
43. E.Shakhnovich, A.Gutin Biophys. Chem. **34**, 187 (1989).
44. V.Pande, A.Grosberg, T.Tanaka *Heteropolymer Freezing and Design: Towards Physical Models of Protein Folding* Reviews of Modern Physics **72**, 259-314 (2000).
45. J.Bryngelson, P.Wolynes Proc. Nat. Acad. Sci. USA **87**, 7524-7528 (1987).
46. E.Shakhnovich, A.Gutin Proc. Nat. Acad. Sci. USA **90**, 7195 (1993).
47. V.Pande, A.Grosberg, T.Tanaka Proc. Nat. Acad. Sci. USA **91**, 12972 (1994).
48. E.Shakhnovich *Modeling protein folding: the beauty and power of simplicity* Folding and Design **1**, 50-52 (1996); E.Shakhnovich *Theoretical studies of protein-folding thermodynamics and kinetics* Current Opinion Structural Biology **7**, 29-40 (1997).
49. J.Onuchic, N.Socci, Z.Luthey-Schulten, P.Wolynes *Protein folding funnels: the nature of the transition state ensemble* Folding and Design **1**, 441-450 (1996).
50. K.Dill, H.Chan, *From Levinthal to pathways to funnels.* Nature: Structural Biology **4**, 10-19 (1997).
51. H.Li, R.Helling, C.Tang, N.Wingreen Science **273**, 666 (1996).